School-by-school calculation

julian bernado

This calculation is at the core of the matchAhead procedure. Let us re-start the notation with treatment school t and control school c then consider treatment units indexed as $T = \{(t, j) : 1 \leq j \leq N_t\}$ with N_t being the size of our treatment school. Then, let the control units be indexed as $C = \{(c, j) : 1 \leq j \leq N_c\}$. Then, we notate our predicted prognostic scores as $\hat{Y}_{ij}$ and our caliper as K. It has up to three steps and may terminate early depending on the results of earlier steps. Regardless of which step it terminated at, it should report three numbers:

1. $B(t, c)$: an aggregate measure of bias along the estimated prognostic scores
2. $E(t, c)$: a measure related to the maximum effective sample size arising
3. $W^{(P)}(t, c)$: Total time in number of seconds required to run the whole school-by-school calculation

the third's calculation should be sufficiently clear, so I'll describe how to attain the first and second numbers depending on steps in the matchAhead process.

Check e_1

We define

$$e_1 = \left| \{(t, j) \in T : |\hat{Y}_{tj} - \hat{Y}_{ck}| < K \text{ for some } 1 \leq k \leq N_c\} \right|$$

where e_1 is the number of treatment units that have some control unit within a caliper along the estimated prognostic score.

If $e_1 = 0$, report no match

If there are no treated units within one caliper of any control unit, then we report an infinite distance for these two schools.

If $e_1 < N_t$, report the blurry look

Let $C^* = \{(c, j) \in C : |\hat{Y}_{cj} - \hat{Y}_{tk}|$ for some $k = 1, ..., N_t\}$ be the set of control units that have some treated units within a caliper. Then, we report:

1. $B(t, c)$: $\left|\text{mean}\left(\{\hat{Y}_{tj}\}_{j=1}^{N_t}\right) - \text{mean}\left(\{\hat{Y}_{cj} : (c, j) \in C^*\}\right)\right|^{-1}$
2. $E(t, c)$: $\frac{N_t}{e_1}$

If $e_1 = N_t$, check e_2

At this point, we now start to think of the school-by-school calculation as a matching problem whose solution is rendered by performing a maximum flow calculation. As such, we imagine a 0-1 caliper distance where the distance between treatment unit j and control unit k is defined as:

$$\mathbf{1}\left(\left|\hat{Y}_{tj} - \hat{Y}_{ck}\right| < K\right)$$

where we're just tracking whether or not the two units are within a caliper of one another. Then, performing a pair matching between T and C along this distance is equivalent to performing a maximum flow calculation on the relevant network. This maximum flow, which we'll notate as M will be a collection of pairs of treated and control units $((t, j), (c, k)) \in T \times C$. Then, we define

$$e_2 = |M|$$

which, since we're doing a pair matching, will be exactly the number of treated units that ended up being matched under this distance.

If $e_2 < N_t$, report the cloudy look

Let $C^\dagger = \{(c, j) : ((t, k), (c, j)) \in M$ for some $k = 1, ..., N_t\}$ be the subset of the controls that was matched under M. Then, we report

1. $B(t, c)$: $\left|\frac{e_2}{\sum_{((t,j),(c,k)) \in M} \hat{Y}_{tj} - \hat{Y}_{ck}}\right|$. This is the average distance within pair matches.
2. $E(t, c)$: $\frac{1}{e_2}$. This is the reciprocal of the number of matches formed as well as the inverse of the effective sample size.

If $e_2 = N_t$, check e_3

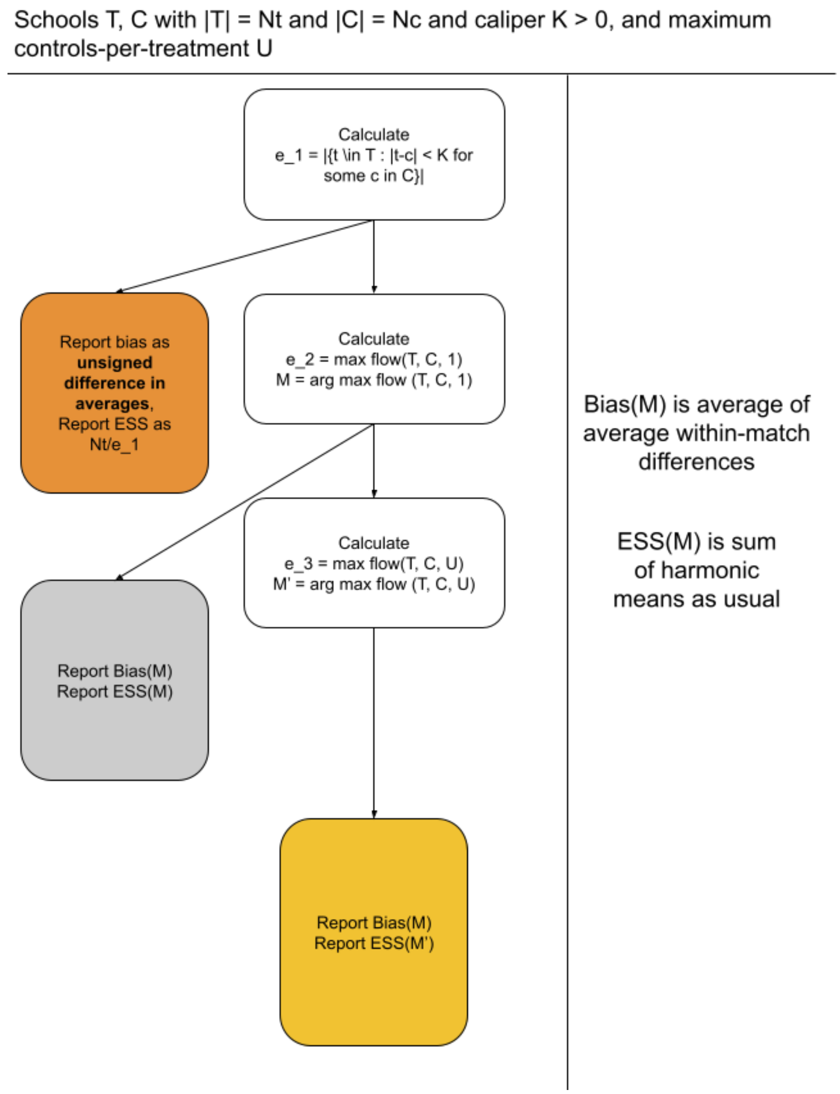
If we have that each of our treated units can be matched in non-overlapping pairs to control units, then we check what would happen if we loosened the maximum number of controls in a match to be U. Then, by using the same 0-1 distance specification but loosening the matching structure, we result in the maximum flow match $M' \subset T \times C$ where a treated unit may now appear up to U times.

Report the clear look

We now report the same things we would in the case where we didn't see e_2, although now accounting for the new matching structure. For a given treatment unit $(t, j) \in T$, let $m(t, j)$ be the number of controls matched to (t, j). Then,

1. $B(t, c)$: $\left| \frac{e_3}{\sum_{j=1}^{N_t} \frac{1}{m(t,j)} \sum_{(c,k):((t,j),(c,k)) \in M'} \left(\hat{Y}_{tj} - \hat{Y}_{ck}\right)} \right|$. This is an adequately-weighted within-group distance.
2. $E(t, c)$: $\left(\sum_{j=1}^{N_t} \frac{2m(t,j)}{1+m(t,j)} \right)^{-1}$. This is the reciprocal of the effective sample size

For a final distance, we report the geometric mean of the two distances: $D_{tc} = \sqrt{B(t,c)E(t,c)}$ along with the total time elapsed W_{tc}. The whole thing is summarized in the following graphic:



Schools T, C with |T| = Nt and |C| = Nc and caliper K > 0, and maximum controls-per-treatment U
Calculate
e_1 = |{t \in T : |t-c| < K for some c in C}|
Report bias as **unsigned difference in averages**, Report ESS as Nt/e_1
Calculate
e_2 = max flow(T, C, 1)
M = arg max flow (T, C, 1)
Report Bias(M)
Report ESS(M)
Calculate
e_3 = max flow(T, C, U)
M' = arg max flow (T, C, U)
Report Bias(M)
Report ESS(M')
Bias(M) is average of average within-match differences
ESS(M) is sum of harmonic means as usual